3. Железнодорожный словарь [Электронный ресурс]. URL: http://rzd.me/inform-block/zhd-slovar/ (дата обращения: 20.09.2017).

4. *Рубцова Ю. В.* Построение корпуса текстов для настройки тонового классификатора // Программные продукты и системы. 2015. № 1 (109). С. 72–78.

5. Документация API социальной сети «ВКонтакте» [Электронный ресурс]. URL: https://vk.com/dev/manuals (дата обращения: 29.09.2017).

УДК 004.891

Е. С. Подоплелова
Научный руководитель: д-р тех. наук, профессор А. Н. Целых
Инженерно-технологическая академия
Южного федерального университета, Таганрог

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ. АЛГОРИТМ C4.5

Аннотация. В этой статье описывается алгоритм интеллектуального анализа данных C4.5. Описывается работа классификатора, его структура, пример. Представлен механизм построения дерева решений. Обосновывается актуальность и особенности рассматриваемого метода.

Ключевые слова: интеллектуальный анализ данных (data mining); экспертная система; алгоритм анализа данных; классификатор; дерево решений; атрибуты.

Интеллектуальный анализ данных (Data Mining) — это неотъемлемая часть любой экспертной системы. Data Mining — это процесс обнаружения в сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных для интерпретации знаний, необходимых для принятия решений в различных сферах. Технологии Data Mining представляют большую ценность для руководителей и аналитиков в их повседневной деятельности. Деловые люди осознали, что с помощью методов Data Mining они могут получить ощутимые преимущества в конкурентной борьбе.

Выбор конкретного алгоритма анализа данных остается за разработчиком на этапе проектирования системы. Существуют уже довольно распространенные, широко используемые варианты. Один из них — C4.5. Основан он на дереве решений. Приобрел популярность благодаря довольно понятному представлению и качественному механизму.

Классификатор. C4.5 строит классификатор в виде дерева решений (рис. 1). Для этого в C4.5 дан набор данных, представляющих вещи, которые уже классифицированы. Классификатор — это инструмент интеллектуального анализа

данных, который берет кучу данных, представляющих вещи, которые мы хотим классифицировать, и пытается предсказать, к какому классу принадлежат новые данные.

Сложно понять сложные деревья решений, потому что информация об одном классе обычно распространяется по всему дереву. C4.5 ввел альтернативный формализм, состоящий из списка правил формы «если A и B и C и … затем класс X», где правила для каждого класса сгруппированы вместе. Случай классифицируется путем нахождения первого правила, условия которого удовлетворяются случаем; если ни одно правило не выполняется, случай присваивается классу по умолчанию.

Построение дерева решений. C4.5 использует два эвристических критерия для ранжирования возможных тестов: усиление информации, которое сводит к минимуму полную энтропию подмножеств {Si} (но сильно смещается к испытаниям с многочисленными результатами) и коэффициент усиления по умолчанию, который делит информационное усиление на информацию обеспечиваемых результатами испытаний. Атрибуты могут быть либо числовыми, либо номинальными, что определяет формат результатов теста. Для числового атрибута A они являются {A ≤ h, A> h}, где пороговое значение h определяется путем сортировки S по значениям A и выбора разделения между последовательными значениями, которые максимизируют указанный выше критерий. Атрибут A с дискретными значениями по умолчанию имеет один результат для каждого значения, но параметр позволяет группировать значения в два или более подмножества с одним результатом для каждого подмножества. Исходное дерево затем обрезается, чтобы избежать переобучения. Алгоритм обрезки основан на пессимистической оценке частоты ошибок, связанной с набором из

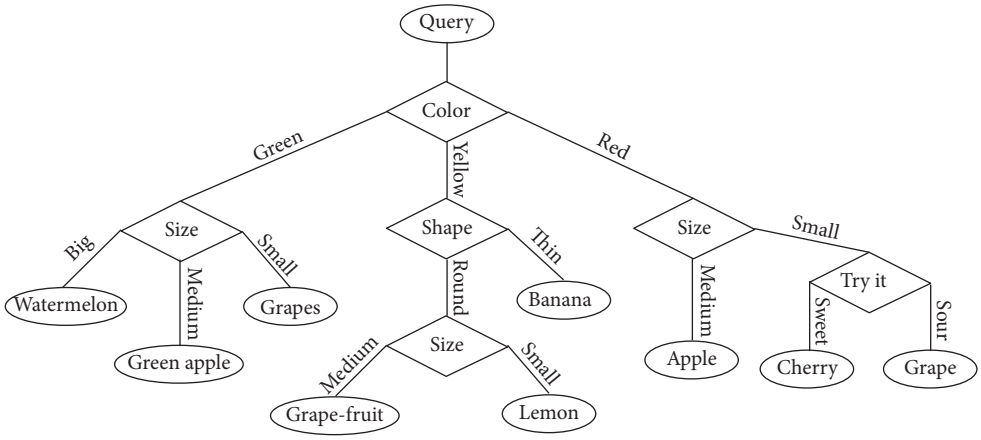

Рис. 1. Классификатор в виде дерева решений

N случаев, E из которых не относятся к наиболее частым классам. Вместо E / N C4.5 определяет верхний предел биномиальной вероятности, когда E-события наблюдались в N испытаниях, используя пользовательскую уверенность, значение по умолчанию которой равно 0,25. Обрезка осуществляется от листьев до корня. Оцененная ошибка на листе с N случаев и ошибок E в N раз превышает пессимистическую частоту ошибок, как указано выше. Для поддерева C4.5 добавляет оценочные ошибки ветвей и сравнивает их с оценочной ошибкой, если поддерево заменяется листом; если последнее не выше первого, поддерево обрезается. Аналогично C4.5 проверяет оценочную ошибку, если поддерево заменяется одной из его ветвей, и когда это кажется полезным, дерево изменяется соответствующим образом. Процесс обрезки завершается за один проход через дерево [1].

Этот подход имеет ряд выгодных особенностей, выделяющих его из списка алгоритмов, основанных на построении дерева решений. Во-первых, C4.5 использует коэффициент усиления информации при генерации дерева решений. Во-вторых, хотя другие системы также включают обрезку, C4.5 использует однопроходный процесс обрезки для смягчения переуплотнения. Обрезка приводит к большим улучшениям в работе. В-третьих, C4.5 может работать как с непрерывными, так и с дискретными данными. Он делает это, задавая диапазоны или пороговые значения для непрерывных данных, тем самым превращая их в дискретные [2].

Главным преимуществом метода является простота интерпретации. Выходные данные понимаются интуитивно. Также у этого метода относительно высокая скорость работы.

Помимо этого, стоит упомянуть алгоритмы:

1. Метод к-средних. Основан на кластеризации. Главным недостатком является его работа только с непрерывными данными.

2. Метод опорных векторов (SVM — Support vector machine). SVM позволяет спроецировать ваши данные в пространство большей размерности. Когда данные спроецированы, SVM определяет лучшую гиперплоскость, которая делит данные на 2 класса [3].

3. Алгоритм Apriori. Ищет ассоциативные правила и применяется по отношению к базам данных, содержащим огромное количество транзакций.

4. AdaBoost. Это алгоритм построения «сильного» классификатора как линейной комбинации:

- AB способен снижать как смещение (например, пни), так и дисперсию (например, деревья) слабых классификаторов;
- AB обладает хорошими свойствами обобщения (максимизирует маржу);
- AB близок к последовательному принятию решений (он создает последовательность постепенно более сложных классификаторов).

Из-за ограничения размера статьи я привела лишь список алгоритмов и существенные особенности, на которые следует обратить внимание [3].

Список литературы

1. Top 10 algorithms in data mining / X. Wu, V. Kumar, J. R. Quinlan et al. Springer-Verlag London Limited. 2007.

2. Top 10 Data Mining Algorithms, Explained [Electronic resource]. URL: https://www.kdnuggets.com/2015/05/top-10-data-mining-algorithms-explained.html (дата обращения: 08.11.2017).

3. *Oza N.*, *Russell S.* Online Bagging and Boosting // Artificial Inteligence and Statistics, 2001.

УДК 004.65

Е. В. Рясов
Научный руководитель: доц. В. Ю. Бердюгин
Южно-Уральский государственный университет, Челябинск

ИСПОЛЬЗОВАНИЕ ВОЗМОЖНОСТЕЙ ИСУБД «CRONOSPRO» ДЛЯ ОРГАНИЗАЦИИ ИНФОРМАЦИОННО-АНАЛИТИЧЕСКОЙ ОБЕСПЕЧЕНИЯ ДЕЯТЕЛЬНОСТИ ПО ЗАЩИТЕ ИСПДн

Аннотация. Деятельность по обеспечению информационной безопасности, как и любая другая организационно-управленческая деятельность, нуждается в информационно-аналитическом обеспечении. Формы и методы организационно-управленческой деятельности применяются в определенной последовательности, цикличности, диктуемой интересами и целями подготовки, принятия и исполнения управленческих решений. Этапы управленческой деятельности имеют логическую связь и образуют в совокупности цикл управленческих действий. В качестве объекта исследования выбрана деятельность по обеспечению защиты информационной системы персональных данных (ИСПДн). Для удовлетворения информационных потребностей, возникающих при защите ИСПДн, предлагается использовать инструментальную систему управления базами данных (ИСУБД) «CronosPro».

Ключевые слова: информационная безопасность; персональные данные; инструментальная система; организационно-управленческая деятельность; информационно-аналитическое обеспечение; ИСУБД «CronosPro».